

An MLE-Based Procedure for Imputing Values for Left-Censored Lognormal Data

Author(s): Philip Villanueva

Affiliation(s): EPA/OPP/HED, Washington, DC

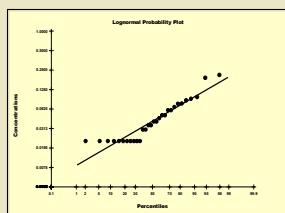
What is “left-censored” data?

Concentrations of chemical compounds, such as pesticides, are found in various environmental and biological media. Often such concentrations are present at levels that cannot be reliably quantified due to limitations of the analytical method. Analytical samples reported below the limit of quantitation (LOQ) are examples of left-censored data. Generally the term “left-censored” refers to data for which only the upper bound is known (i.e. known to be below some threshold value).

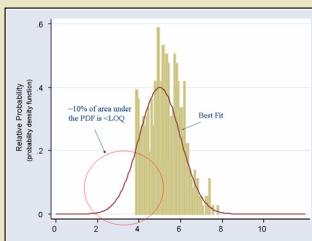


Result of a Simple Imputation Procedure

A commonly used imputation procedure is to report below LOQ measurements as some multiple (e.g., 1/2 or 1/10) of the LOQ. Such simple imputation procedures will result in biased estimates when calculating summary statistics, inadequate statistical intervals, or potentially invalid conclusions when performing parametric statistical tests. Probability plots characterized by a “flat” lower portion are useful identifying data sets with such imputed values.



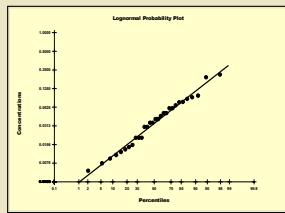
What is maximum likelihood estimation?



Maximum likelihood estimation (MLE) involves selecting the distribution amongst a family of distributions (e.g. lognormal) that best fits the observed dataset. Each distribution within a family is specified by a set of parameters that can be used to calculate various statistics of interest, such as the mean and standard deviation. An iterative procedure is employed to select the combination of parameters that maximizes the likelihood of obtaining the observed dataset. The distribution specified by the parameters that maximize the likelihood function is termed the “best fit.”

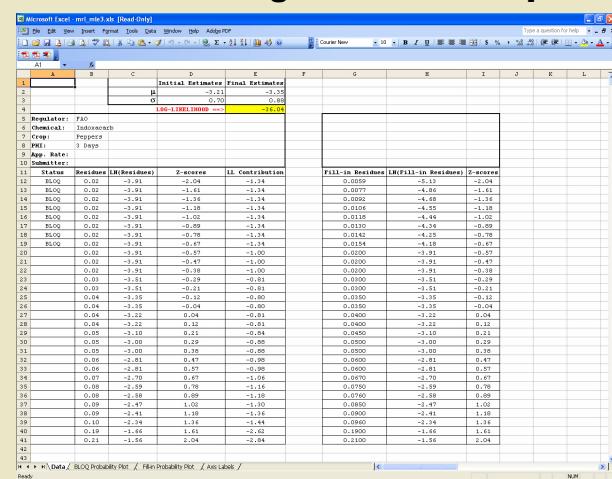
Result of MLE-Based Imputation Procedure

The parameters estimated via MLE can be used to calculate values for the left-censored samples that are consistent with the fitted distribution. These imputed values provide more realistic values than those produced by simple imputation procedures. One of the advantages of using such an imputation procedure is that MLE techniques incorporate information regarding the proportion of the dataset that is reported below the LOQ (i.e. the censoring threshold) in addition to the values of the quantifiable concentrations. MLE requires that a distribution type, such as lognormal, be specified. Environmental measurements generated by a common process (e.g. pesticide applied to a crop) are often thought to be lognormally distributed.



A Simple Spreadsheet for Calculating MLE-Based Imputed Values for Lognormal Data

The Office of Pesticide Programs' (OPP) Health Effects Division (HED) has developed an Excel spreadsheet that provides the MLE-based lognormal parameter estimates for left-censored datasets utilizing Excel's “Solver” optimization tool. Based on the MLE parameter estimates, “fill-in” (i.e. imputed) values are calculated for the left-censored data that are consistent with a lognormal distribution. Additionally, lognormal probability plots are automatically generated to compare the original dataset to the MLE-supplemented dataset. The supplemented dataset can then be used to calculate summary statistics and statistical intervals, perform parametric statistical tests, or inputted directly into a dietary risk assessment.



Although the MLE spreadsheet was originally created to provide more realistic imputation values for pesticide field trial datasets containing samples reported below LOQ, the spreadsheet could be used for any left-censored dataset thought to be lognormally distributed, such as biomonitoring data and compliance monitoring data. Contact Philip Villanueva (villanueva.philip@epa.gov) to obtain a copy of the spreadsheet.



epaScienceforum
Your Health • Your Environment • Your Future